DOCUMENT RESUME

ED 189 098 TM 800 116

AUTHOR Marco, Gary L.: And Others
TITLE A Test of the Adequacy of Curvilinear Score Equating Models.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.: Educational Testing Service, Princeton, N.J.
PUB DATE 28 Jun 79
NOTE 51p.: Paper presented at the Computerized Adaptive Testing Conference (Minneapolis, MN, June 27-30, 1979).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Aptitude Tests: *Equated Scores: Error of Measurement: High Schools: *Latent Trait Theory: Multiple Choice Tests: Sampling: Statistical Bias: True Scores: Verbal Tests
IDENTIFIERS *Classical Test Theory: *Equipercentile Equating: Scholastic Aptitude Test

ABSTRACT

Data from the verbal portion of the College Entrance Examination Board Scholastic Aptitude Tests were used in an experimental test of the accuracy of equating for a variety of models in three categories: linear equating, equipercentile equating, and item characteristic curve equating. The models were tested for both mean squared error and bias. These methods were used to equate a test to itself, to another parallel test, and to another test that differed in difficulty. Samples used in the equating were designed to be either random or to be different from each other in ability. Equating was based on subsamples of 1577 cases selected from samples of 4731 cases. Criterion equatings used all the cases in each sample and used either item characteristic curve equipercentile methods or direct observed score equipercentile methods. A number of conclusions were developed indicating that the most effective type of equating depended on the characteristics of the tests and the samples being analyzed. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Educational Testing Ser TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DRAFT COPY. Not for Circulation without the express permission of the author

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED189098

TM800 116

A Test of the Adequacy of Curvilinear Score Equating Models

Gary L. Marco, Nancy S. Petersen, and E. Elizabeth Stewart

Educational Testing Service, Princeton, New Jersey

In many common testing situations it is necessary to compare the test scores of examinees who have taken different forms of a test. In practice, two forms of a test cannot be expected to be of exactly equal difficulty for examinees at all ability levels. Therefore, a comparison of raw scores on two forms of a test will be unfair to the examinees who have taken the more difficult form. Statistical procedures that have been developed to deal with this problem are referred to as equating methods.

In an ideal psychometric world, tests on which scores need to be equated would be parallel in all important respects; an anchor test, if used, would be parallel to the total tests; and random samples on which to base the equating would always be available. In actual testing practice, however, scores must sometimes be equated under less-than-optimum conditions. This study is the first part of a fuller study, the purpose of which is to examine the adequacy of score equating models when certain sample and

---

Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June 28, 1979. The authors are listed alphabetically. The study was funded jointly by the College Entrance Examination Board and Educational Testing Service. The authors gratefully acknowledge the advice and assistance provided by William H. Angoff, Rebecca Bullock, Samuel A. Livingston, Frederic M. Lord, Ruth C. Mroczka, Joanne S. Peng, William B. Schrader, Ledyard R Tucker, Marilyn S. Wingersky, and R L Wood.

test characteristics are systematically varied. The emphasis in this part of the study is on curvilinear models. The second part focuses on linear models. This study is more comprehensive than previous studies of equating models (e.g., Levine, 1955; Tucker, 1974; Rentz and Bashaw, 1975; Slinde and Linn, 1977, 1978) in that it includes a greater variety of equipercentile, linear, and ICC models and investigates equatings based on dissimilar samples as well as on random samples.

EQUATING MODELS

An equating method is an empirical procedure for determining a transformation to be applied to the scores on one of two forms of a test. Its purpose is (ideally) to transform the scores in such a way that it makes no difference to the examinee which form of the test he or she takes. This ideal can be reached only if the two forms of the test measure exactly the same latent trait (ability or skill) and yield scores that are equally reliable and if the equating transformation is invertible.

Because an equating method is an empirical procedure, it involves a design for data collection and a rule for determining the transformation. There are three basic designs for data collection and three general rules for determining the transformation. Any of the three designs for data collection can be used with any of the three transformation rules.

The three designs for data collection are the single-group method, the equivalent-group method, and the anchor-test method (Lord, 1975). All of the equating procedures used in this study assume that the data were collected using the anchor-test method. An anchor-test design requires administering one form of a total test to one group of examinees, a second form to a second group of examinees, and a common anchor test

to both groups. The anchor test can be either internal or external to the tests to be equated. The anchor test is used to reduce equating bias resulting from differences in ability between the two groups.

The three general rules for determining the transformation can be described as follows:

A. Equipercentile equating. Choose a transformation such that scores from the two tests will be equated if they correspond to the same percentile rank in some group of examinees.

B. Linear equating. Choose a linear transformation such that scores from the two tests will be equated if they correspond to the same number of standard deviations from the mean in some group of examinees.

C. Item characteristic curve (ICC) equating. Choose a transformation such that true scores from the two tests will be equated if they correspond to the same estimated level of the latent trait underlying both tests.

All three types of equating were represented in this study.

Equipercentile and linear methods using anchor-test data can be further classified as to whether the equating is done directly or indirectly by frequency estimation. In direct equipercentile equating, scores on each test and on the anchor test are first equated separately within each group. Then scores on the two tests to be equated are said to be equivalent if they correspond to the same score on the anchor test. Frequency estimation (Angoff, 1971, pp. 581-582), on the other hand, makes use of the combined distribution of

scores on the anchor test. The score distributions of the tests to be equated are estimated for the combined group of examinees, and these estimated distributions are then used as if they had been observed from a single-group design. The resulting marginal distributions on the two forms are used in the case of equipercentile equating; the resulting estimated means and standard deviations on the two forms are used in the case of linear equating.

Many of the linear methods require error variance estimates. The three methods of estimating error variances that were used in this study were Angoff's method (1953), which uses anchor-test data, Feldt's method (1975), which uses part-test data, and coefficient alpha, which uses item-response data.

One of the ICC procedures utilized the ICC parameters (item difficulties) from the 1-parameter logistic test model; the other utilized the ICC parameters from the 3-parameter logistic test model. The computer program LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) was used to estimate item parameters and examinee abilities. For both models, true formula scores (R-W/4) were equated by calculating the true formula scores on each test form corresponding to selected ability levels (Lord, 1975) and interpolating as necessary. Since for either model there is a functional relationship between ability and true score, the true formula score is readily computed by the formula $R_g - (N - R_g)/(A - 1)$, where R is the true number right score at ability g computed by summing the item proportions correct under the model, N is the number of test items, and A is the number of response options for the items in the test (5-choices in all cases).

The various equating models used in this study are described briefly in Table 1. The models are categorized by whether the procedure results in a linear or curvilinear transformation between observed or true scores.

The table provides references and information as to the major assumption underlying each model, the kind of data required, and whether specific error variance estimates are needed. If the codes for the model differ for an external and an internal anchor test, then the formulas for computing the transformation parameters differ in the two cases. A total of 40 linear (including four based on the marginal means and standard deviations resulting from frequency estimation), two equipercentile, and two ICC equating models were used in the study.

STUDY DESIGN

ETS computer files for two national administrations, April 1975 and November 1975, of the verbal portion of the Scholastic Aptitude Test (SAT-Verbal) served as the source of the data for the study. Fourteen pairs of total test scores were equated on the basis of the data from each of the two administrations. Records of test scores, responses to test items, and responses to a Student Descriptive Questionnaire (SDQ) were accessed to construct samples having specified characteristics, to extract or calculate selected scores for a number of special-purpose total tests and anchor tests, and to compute item statistics and other data needed for the various equating models.

Equating Design

The total-test--anchor-test combinations used in the various equatings for each of the two administrations can be classified into five categories, referred to here as test variations. (There were 11 test variations used in the full study of which this study is a part.) The test variations studied are outlined in Table 2.

Equating a Test to Itself

In this part of the study, a single form of the operational verbal portion of the SAT was treated as if it represented two different forms; that is, it was to be equated to itself. (This type of design was first used by Levine, 1955.) This part of the study was designed to investigate the effects of varying (a) the relative difficulty levels of the total test and the anchor test and (b) the degree of similarity between the two samples on which the equating operations were based.

SAT-Verbal was equated to itself through two anchor tests that, like the total test, were of medium difficulty. One was external (Test Variation 1), and one was internal (Test Variation 4). It was also equated to itself through two internal anchor tests that differed from it in difficulty (Test Variation 5). One of the internal anchor tests was easier than SAT-Verbal and the other, harder.

For each administration, a single pair of random samples and a single pair of dissimilar samples were used for all equatings of SAT-Verbal to itself. The dissimilar samples were ones that, by virtue of the sample selection procedure, were expected to be of middle and high verbal ability, respectively.

Equating Tests of Different Difficulty

In this part of the study, three total tests were constructed for each administration from a pool comprising the operational SAT-Verbal items and the items in a nonoperational section of verbal material. The purpose of this part of the study was to examine the effects of varying (a) the relative difficulty levels of the two total tests on which scores were to be equated and (b) the degree of similarity between the two equating samples.

Pairs of total tests constructed to differ in difficulty were equated through an internal anchor test of medium difficulty. The random and dissimilar samples used in equating SAT-Verbal to itself were used in these equatings also, along with an additional sample expected to be of low verbal ability.

In Test Variation 8 an easy test was equated to a medium-difficulty test, and the medium-difficulty test was equated to a hard test. In Test Variation 9 the easy test was equated to the hard test. For equatings based on dissimilar samples, data from the low-ability sample was used for the easy test; from the middle-ability sample, for the medium-difficulty test; and from the high-ability sample, for the hard test.

Tests

Several scores calculated as part of the normal processing for SAT administrations were included in the study. Included also were scores computed on a number of tests constructed retrospectively solely for use in the study.

The general approach followed in constructing the special-purpose tests entailed (a) developing content and statistical specifications for all special-purpose tests, (b) identifying sets of items in accordance with those specifications, and (c) identifying subsets of items on which separate scores were to be obtained for use in calculating one of the three sets of reliability estimates required for the equating analyses.

Test content was specified only in terms of distributions of item types, although more detailed specifications are followed in developing operational forms of SAT-Verbal. SAT-Verbal is composed of three types of discrete 5-choice items--antonyms (25 items), analogies (20 items), and sentence completions (15 items)--and of five reading comprehension

passages each of which is followed by five 5-choice items based on the passage.

Statistical specifications were stated in terms of the item statistics conventionally used in the development of ETS tests (described by Angoff and Dyer, 1971, pp. 9-10). The equated delta (Δ_e) served as the index of item difficulty and the biserial correlation (r_b) of the item score with the total score on the operational test in which the item appeared, as the index of item discrimination. The statistic Δ_e is an estimate of the difficulty of the item for a standard reference group. It ranges in value from about 6 (very easy) to about 18 (very hard). If a test composed of 5-choice items were of middle difficulty for the reference group, its mean Δ_e would be 12.0.

The item statistics used to construct the tests were taken from the results of item analyses routinely conducted for each new form of the SAT. The analyses were based on systematic samples of approximately 1,700 to 2,000 examinees each.

After all full-length total tests and anchor tests had been identified, part tests for use in reliability estimation (Feldt, 1975) were created. Each part test was parallel, except for length, to the full-length test from which it was derived.

Tests Used in Equating a Test to Itself

For each administration scores were available on an external anchor test (a nonoperational section of verbal material) similar in content to SAT-Verbal. The external anchor tests each contained equal numbers of items of the four types included in the SAT-Verbal. The difficulties of the external anchor tests were not subject to experimentation. The mean difficulties of the external anchor tests were within about half a

Δ_e point of the mean for SAT-Verbal, but both the standard deviations of the Δ_e's (σ_{Δ_e}) and the mean r_b's tended to be somewhat lower for the anchor tests.

For each administration, three internal anchor tests for equating SAT-Verbal to itself were specially constructed for the study from the pool of 85 operational items in each form. The internal anchor tests were constructed to be similar in content to SAT-Verbal but to vary with regard to each other in mean difficulty. Internal anchor tests constructed from the April 1975 item pool each contained 10 antonym, 6 analogy, 8 sentence completion, and 10 reading comprehension items. The number of items in these respective categories from the November 1975 item pool were 10, 8, 8, and 10. For the medium-difficulty anchor tests, $\overline{\Delta}_e$, σ_{Δ_e}, and $\overline{r}_b$ were made to match the corresponding values for SAT-Verbal as closely as possible, given the prescribed content distribution. The σ_{Δ_e}'s for the easy and hard anchor tests were, of necessity, smaller than those for SAT-Verbal and the medium-difficulty anchor test.

The item summary statistics and identification codes for the total tests and anchor tests used in equating a test to itself are given in Table 3.

Tests Used in Equating Different Tests

For each administration, an expanded item pool consisting of the 85 SAT-Verbal items and the 40 items in the verbal external anchor test was used for creating three total tests. The tests within each set were systematically different in average difficulty but were similar in content and equal in length. The total tests constructed from the April 1975 and from the November 1975 item pools each contained 15 antonym, 13 analogy, 11 sentence completion, and 15 reading comprehension items.

For each administration, a 20-item internal anchor test of medium difficulty, similar in content to SAT-Verbal, was constructed for use in equating the special-purpose total tests. The internal anchor tests constructed from the April 1975 and from the November 1975 item pools each contained 6 antonym, 5 analogy, 4 sentence completion, and 5 reading comprehension items.

The item summary statistics and identification codes for the total tests and anchor tests used in equating different tests are given in Table 3.

Samples

Two base samples were used for the study, one each from the April 1975 and November 1975 Saturday administrations of the SAT. The April base sample (No. 32) was selected from those candidates who took verbal equating test fm; the November base sample (No. 44), from those who took verbal equating test fg. (Six base samples were used in the full study.)

Each base sample consisted of 4,731 cases, from which five subsamples of 1,577 cases each were created in two different ways. Two nonoverlapping subsamples ("random" samples) were selected by use of an IBM recursive random number generator. Three nonoverlapping subsamples ("dissimilar" samples) were selected by an algorithm designed to yield samples dissimilar in mean verbal ability. Two variables from the Student Descriptive Questionnaire, level of educational aspiration and amount of high-school foreign language training, were used to select the dissimilar samples. These variables were known from prior information to have a high relationship with SAT-Verbal scores.

Thus, a total of 10 subsamples of 1,577 cases each, five for each of the two base samples, were used in the study. Tables 4 and 5 give the

summary statistics for the total tests and anchor tests used in the various equating analyses.

Evaluative Procedures

Discrepancy Indices

For each raw score x there is a corresponding criterion score t and an estimated criterion score t' derived from a specific equating model. The smaller the difference d between t' and t, the smaller the equating error and the more appropriate the equating model.

The standardized weighted mean square difference and squared bias were selected as the most useful summary indices for evaluating the effectiveness of the various models. The weighted mean square difference gives the greatest weight to those values of x which are most likely to occur and is consistent with what is used to represent total error in the statistical literature. The indices were standardized (expressed as a proportion of the criterion standard deviation) so that results could be compared across equating situations as well as across equating models.

The standardized weighted mean square difference or total error is equal to the variance of the difference plus the squared bias, that is,

$$\sum_j f_j d_j^2/ns_t^2 = \sum_j f_j (d_j-\bar{d})^2/ns_t^2 + \bar{d}^2/s_t^2 \quad \text{or}$$

(Total Error) = (Variance of Difference) + (Squared Bias),

where $d_j = (t'_j - t_j)$, t'_j is the estimated criterion score for raw score x_j, t_j is the criterion score for x_j, $\bar{d} = \sum_j f_j d_j/n$, s_t is the standard deviation of the criterion scores t_j, f_j is the frequency of x_j, $n = \sum_j f_j$, and the summation was over that range of x for which extrapolation was unnecessary for any of the models studied. If the ratio of the squared bias to the total error is 1, then the criterion line and the conversion

line are parallel. If the difference is less than 1, then there is an interaction between the model and the criterion.

Criterion Equatings

In the case where a test was equated to itself, the criterion for the various equatings was the test score itself. The new and old forms were treated as different tests when in reality they were one and the same test. The ideal equating would reproduce the score on the old form exactly; that is, the conversions from raw to scaled scores would be the same for the new form as for the old form. The criterion was not so simply established in the case where a test was equated to a different test. In these instances, it was necessary to calculate "true" conversions by equating the tests in as ideal a manner as possible.

The criterion equatings were accomplished using data from all of the cases in the two base samples (No. 32 and No. 44) from which subsamples were selected. Since all 4,731 cases in each base sample had scores on the tests being equated, it was possible to equate the scores using a single sample, an ideal equating situation. The scores could be linked directly without involving an anchor test.

Two equating methods were used to establish the two criteria against which to compare the results of the experimental equatings: equipercentile equating of estimated true scores derived from the 3-parameter logistic test model, the "ICC equipercentile criterion," and equipercentile equating of observed scores, the "direct equipercentile criterion." Although the criterion equatings using estimated true scores were based on item characteristic curve methodology, the method was different from that used in the experimental equatings. Nevertheless, the ICC equipercentile criterion could be biased in favor of ICC equating methods. Thus, the

direct equipercentile criterion was also used. This criterion might be biased in favor of equipercentile, but not ICC, methods. For the study of the curvilinear methods reported here, both methods were appropriate, since the total tests were equal in length and yielded scores with nearly equal reliabilities. (True-score equating will generally yield better conversions than observed-score equating when the new and old form scores have unequal reliabilities.)

To determine the ICC equipercentile criterion, the 3-parameter logistic test model was applied separately to Reading (reading comprehension and sentence completion) and Vocabulary (antonym and analogy) items, so that unidimensionality did not have to be assumed across all item types. The following scores were equated:

Base Sample No. 32		Base Sample No. 44	
New Form	Old Form	New Form	Old Form
FT2DE	FT2DM	FT4DE	FT4DM
FT2DE	FT2DH	FT4DE	FT4DH
FT2DM	FT2DH	FT4DM	FT4DH

The following steps were required to accomplish the 3-parameter logistic item characteristic curve criterion equatings:

1. Use LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) to calculate item parameter estimates and examinee ability estimates separately for each set of relatively homogeneous items (Reading and Vocabulary).

2. For each of the three scores, calculate for each examinee an estimated true raw (R - W/4) score across the item types represented in each score. The true raw score (R_a) was estimated as follows:

$$R_a = \sum_i P_i(\hat{\theta}_{ia}) - [\sum Q_i(\hat{\theta}_{ia})]/4,$$

where $\hat{\theta}_{ia}$ is the estimated ability of examinee a on the item type represented by item i (if item i is a Reading item, then $\hat{\theta}_{ia}$ is the estimate of the examinee's reading ability, etc.), P_i is the probability of

examinee a answering item i correctly (as calculated from the 3-parameter logistic test model), $Q_i = 1 - P_i$, and 1/4 is the correction factor for guessing for 5-choice multiple-choice items. Each summation was over only those items to which examinee a actually responded.

3. For each of the three pairs of scores, equate the scores directly by the equipercentile method. Raw and scaled score equivalents were generated for each integral score on the new form for which it was possible to establish a conversion. (The true raw scores did not usually extend over the possible score range, thus making it impossible to establish a conversion for some scores.)

RESULTS AND DISCUSSION

Selected results are presented in Tables 6 through 11 and Figures 1 through 6. Following an explanation of the tables and figures, salient features of the results are discussed, some limitations on their interpretation are suggested, and several conclusions are presented.

Comparisons of Equating Models

Tables 6 through 11 and Figures 1 through 6 summarize discrepancies between the results of the experimental equatings and the criterion equatings. The discrepancies are stated as mean square error and squared bias. To make the results comparable for different tests, the discrepancies are expressed on a scale on which the standard deviation of the criterion scores would be 100 for a standard reference group. The discrepancy indices for Tables 8 and 10 and Figures 3 and 5 were calculated in relation to criterion equatings in which the 3-parameter logistic test model was used, the "ICC equipercentile criterion." The discrepancy indices for Tables 9 and 11 and Figures 4 and 6 were calculated in relation to criterion

equatings in which the equipercentile equating method was used, the "direct equipercentile criterion."

The test variations represented in Tables 6 through 11 are characterized in Table 2. The codes in the column labeled Ref. Table denote reference tables of source data that do not appear in this report. The statistical characteristics of the tests on which the total-test scores and the anchor-test scores were based are described in Table 3. The first three digits of the numeric codes in the column labeled Sample Number identify the subsamples on which the equatings were based (see Tables 4 and 5 for sample statistics).

The column labeled Best Lin identifies the linear model that had the smallest mean square error under the condition specified. The linear models considered are enumerated in Sections I (observed score models) and III (true score models) of Table 1, which also gives the identification codes for all models. The models identified in Tables 6 through 11 as Equi% (Dir), Equi% (FE), 3-Par ICC, and 1-Par ICC correspond, respectively, to sections IIB, IlA, IVA2, and IVA1 of Table 1.

Figures 1 through 6 correspond sequentially to Tables 6 through 11. It is essential to note that the size of the scale units represented on the y-axis is different in different figures. The scale on the left side of each figure is in terms of standard deviation units and is the square root of the linear scale of mean square error shown on the right side of the figure.

Contiguous bars represent the results for one model or, in the case of BEST LIN, for the best of a set of models. The results for random samples are shown in the top half of the figure and the results for dissimilar samples, in the bottom half. The outer bars represent total error and the inner shaded bars, bias.

The bars in each set, from left to right, appear in the successive rows in the comparable area of the corresponding table. For example, the first bar in each set in Figure 1 represents the results obtained with an internal anchor test for a pair of April 1975 samples, either random or of middle and high ability, respectively; the second bar, an external anchor test for the same April samples; the third bar, an internal anchor test for a pair of November 1975 samples; and the fourth bar, an external anchor test for the same November samples.

Equating a Test to Itself

Test Variations 1 and 4. For both similar and dissimilar samples, the best linear model had the smallest amount of total error across replications, followed by the 1-parameter ICC and the 3-parameter ICC models. The direct equipercentile and frequency estimation equipercentile methods had relatively more error. Surprisingly, the total error for dissimilar samples was very similar to the total error for random samples, implying perhaps that, if the anchor test is nearly parallel to the total tests, the differences between samples is not so important. The only method showing noticeably more error for dissimilar samples was the frequency estimation equipercentile method. In this instance, bias accounted for a large proportion of the error. For dissimilar samples, the equatings through an external anchor had noticeably more error for all but the 1-parameter ICC model.

Test Variation 5. For random samples, the total errors for Test Variation 5 were similar to the total errors for Test Variations 1 and 4. The fact that the total error for the best linear model was small for random samples suggests that the curvilinear relation of the anchor test and the total test had little effect on the equating.

The total error for the best linear model was substantially larger when dissimilar samples were used. The ICC models were noticeably superior to the other models in this case, suggesting that these models are relatively robust when the anchor test is different in difficulty from the total test and the samples differ in ability.

Equating a Test to a Different Test

Test Variations 8 and 9. The introduction of some curvilinearity in the relation between the scores of the tests being equated (due to differences in difficulty) resulted in a very large total error for the best linear model (as expected). For the most part, sample variation seemed to have little effect on total error for the curvilinear models. All of the curvilinear models had noticeably less error than the best linear model, but the 1-parameter ICC model had substantially more error than the other curvilinear models. For most equatings, the model with the smallest total error was the 3-parameter ICC model. For the most part, when greater differences in total test difficulty were introduced (greater curvilinearity), the total error tended to increase substantially for all models, becoming exceedingly large for the best linear model.

Criterion Bias

Equating a Test to a Different Test

The total error and bias for the two criteria for Test Variations 8 and 9 can be compared in Figures 3 and 4 and in Figures 5 and 6. It may be noted that for random samples the experimental equipercentile equatings had less error than the 3-parameter ICC model when the equipercentile observed-score criterion was used, suggesting that the equipercentile criterion equatings based on true scores may be biased in favor of the

3-parameter ICC model, just as the equipercentile criterion equatings based on observed scores seem to be biased in favor of the equipercentile models. Interestingly enough, the 3-parameter ICC model has less total error than the 1-parameter ICC model regardless of which criterion was used, although the difference between the models was less for the direct equipercentile criterion.

The criterion bias is less obvious in the case of dissimilar samples; in fact, the rank ordering of the models in terms of total error was not affected by the choice of criterion. The 3-parameter ICC model had the smallest error under both criteria, but the size of the error decreased for the equipercentile and 1-parameter ICC models and increased for the 3-parameter ICC model when the direct equipercentile criterion was used.

Equating a Test to Itself

The criterion would seem to be well established when a test is equated to itself. However, it is possible that the criterion in this case is biased in favor of a model that comes closest to fixing all of the item parameters at the same values, in this case the 1-parameter ICC model.

It is easily seen that if the a's, b's, and c's in the 3-parameter logistic test model are fixed at some constant values, the true raw scores for the old and new forms corresponding to a given ability level have to be the same, since the probability of getting a particular item correct is then a function of only the item parameters and ability.

In the 1-parameter logistic test model, the c is fixed at 0 and the a at 1 for all items. Since only the b's are estimated, one might expect the 1-parameter ICC model to be more likely than the 3-parameter ICC model to yield the appropriate conversions. Is it of any consequence, then, if we find that the 1-parameter ICC model seems to be superior to the 3-parameter

ICC model in equating a test to itself? What we really want to know is which of the models works best when a test is equated to a different test, particularly when a test is equated to a parallel test.

It would seem then that there may be a natural bias in favor of the 1-parameter ICC model when a test is equated to itself, but this probably does not affect the other models.

Conclusions

Several conclusions can be drawn from the study. They should be considered tentative because of possible criterion bias, which has been only partially controlled; because it was not possible to study all test variations in the full design, in particular, the variations in which parallel but not identical tests were equated; and because the results showed occasional inconsistencies that have not yet been explained.

1. When a test is equated to itself (or, to generalize, to a test like itself) through a parallel anchor test, a linear model yields very good results regardless of the type of samples used. However, whether any particular linear model consistently gives satisfactory results requires further study. The best linear model for a particular test variation and type of sample was used here.

2. Curvilinear models give results nearly comparable to those of the best linear model when a test is equated to a test like itself through an internal anchor. When an external anchor is used, of the curvilinear models, the ICC models (particularly the 1-parameter model) give relatively better results. However, the criterion may be biased in favor of the 1-parameter model.

3. The types of samples have a relatively small and unsystematic effect on the quality of the equating results if the anchor test is

similar in content and difficulty to the total tests. The one exception is the frequency estimation procedure, which seems not to perform well when an external anchor and dissimilar samples are used.

4. The equatings involving an internal anchor have less total error than comparable equatings with an external anchor. Whether this inconsistency would obtain when a test is equated to a test of different difficulty needs to be studied. The possibility that the criterion is biased in favor of an internal anchor also needs investigation. However, it may simply be due to the fact that the external anchor tests were not quite as similar to the total tests in content and statistical characteristics as were the internal anchor tests.

5. When a test is equated to a test like itself through an easy or hard anchor test, with random samples, all of the models have a small mean square error. When dissimilar samples are used, however, the ICC models give clearly superior results.

6. When total tests differ considerably in difficulty, linear models yield unsatisfactory results in that the mean square error becomes very large, but they tend to yield better estimates of the mean than the curvilinear models. The 1-parameter ICC model and the frequency estimation method also give unacceptable results in many instances. This result is consistent with the Slinde and Linn (1978) findings that the Rasch model yielded poor results for vertical equating.

7. The 3-parameter ICC model is the best equating model when total tests of unequal difficulty are equated through a medium-difficulty anchor test with dissimilar samples. It would appear that the 3-parameter model is the equating model most likely to yield acceptable results under unusual or extreme conditions.

This study, though comprehensive in its coverage, is limited in that SAT-Verbal items made up the item pool. These items are known to be relatively homogeneous and somewhat difficult for the current test-taking population. Similar studies are needed in situations where the content of the anchor test is allowed to depart by varying degrees from the content of the total tests and where the test-taking population is from a different age group.

REFERENCES

Angoff, W. H. Test reliability and effective test length. Psychometrika, 1953, 18, 1-14.

Angoff, W. H. Scales, norms, and equivalent scores. In Robert L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Angoff, W. H., & Dyer, H. S. The Admissions Testing Program. In William H. Angoff (Ed.), The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board, 1971.

Birch, M. W. A note on the maximum likelihood estimation of a linear structural relationship. Journal of the American Statistical Association, 1964, 54, 1175-1178.

Feldt, L. S. Estimation of the reliability of a test divided into two parts of unequal length. Psychometrika, 1975, 40, 557-561.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, N.J.: Educational Testing Service, 1955.

Lord, F. M. A survey of equating methods based on item characteristic curve theory (RB-75-13). Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M. Personal communication, February 24, 1976.

Potthoff, R. F. Equating of grades or scores on the basis of a common battery of measurements. In P. R. Krishnaiah, Multivariate Analysis. Academic Press, 1966.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model (Vols. I and II). Athens, Ga.: Educational Research Laboratory, College of Education, University of Georgia, September 1975.

Slinde, J. A., & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Tucker, L. R. Personal communication, August 7, 1974.

Wood, R L, & Lord, F. M. A user's guide to LOGIST (RM-76-4). Princeton, N.J.: Educational Testing Service, 1976.

Wood, R L, Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton, N.J.: Educational Testing Service, 1976.

Table 1

Description of Equating Models

Major Assumption	Data	Error Variances Known: External Anchor	Error Variances Known: Internal Anchor	Codes: External Anchor	Codes: Internal Anchor	Model	Reference
I. Linear Observed Score Models							
A. Constancy of Regression							
1. X on v and Y on v							
a. Curvilinear	Bivariate Distributions	None	None	FE-O	FE-O	Frequency Estimation	Angoff, 1979 pp. 581-582
b. Linear	Cross-Products Matrices	None	None	TI	TI	Tucker I	Angoff, 1971 pp. 576-577
2. X on v' and Y on v' are linear	Cross-Products Matrices	v_a, v_b	v_a, v_b	T2O-E[a]	T2O-I[b]	Tucker 2 Observed	Tucker, 1974
3. V on x' and V on y' are linear	Cross-Products Matrices	x_a, y_b	x_a, v_a, y_b, v_b	T3O-E[a]	T3O-I[b]	Tucker 3 Observed	Tucker, 1974
B. X and V and Y and V Congeneric and Constancy of Regression of X' on v' and of Y' and v'	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	LER[a]	LER[b]	Levine Equally Reliable	Levine, 1955
C. Distributional Assumptions							
1. X|v and Y|v are normal	Cross-Products Matrices	None	None	PAC[h]	PAC[i]	Potthoff A	Potthoff, 1966
2. V|x and V|y are normal	Cross-Products Matrices	None	None	PB	PB[f,g]	Potthoff B	Potthoff, 1966
3. X and V and Y and V are bivariate normal	Cross-Products Matrices	None	None	PAC[h]	PAC[i]	Potthoff C	Potthoff, 1966
4. X and V and Y and V are more general than bivariate normal to allow for group differences	Cross-Products Matrices	None	None	PD	PD	Potthoff D	Potthoff, 1966

Table 1 (Continued)

Major Assumption	Data	Error Variances Known: External Anchor	Error Variances Known: Internal Anchor	Codes: External Anchor	Codes: Internal Anchor	Model	Reference
II. Curvilinear Observed Score Models							
A. Constancy of Regression—X on v and Y on v are curvilinear	Bivariate Distributions	None	None	FE-EQ	FE-EQ	Frequency Estimation	Angoff, 1971, pp. 581-582
B. No explicit assumptions	Marginal Distributions	None	None	EQ	EQ	Equipercentile	
III. Linear True Score Models							
A. Constancy of Regression							
1. X′ on v′ and Y′ on v′							
a. Curvilinear	Bivariate Distributions	x_a, y_b	x_a, y_b	FE-T	FE-T	Frequency Estimation	Angoff, 1971, pp. 581-582
b. Linear	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	T2T-E[c]	T2T-I[d]	Tucker 2 True	Tucker, 1974
2. V′ on x′ and V′ on y′ are linear	Cross-Products Matrices	x_a, y_b	x_a, v_a, y_b, v_b	T3T-E[c]	T3T-I[d]	Tucker 3 True	Tucker, 1974
B. X and V and Y and V Congeneric							
1. Constancy of regression of X′ on v′ and of Y′ on v′	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	LUR[e]	LUR[f]	Levine Unequally Reliable	Levine, 1955
2. Observed scores on test form are linear function of observed scores on ideal test. True scores on ideal test are multiplicative function of true scores on anchor test.	Cross-Products Matrices	x_a, v_a, y_b, v_b	None	TML-E[e]	TML-E[f,g]	Tucker Modified Levine	Tucker, 1974

Table 1 (Continued)

Major Assumption	Data	Error Variances Known: External Anchor	Error Variances Known: Internal Anchor	Codes: External Anchor	Codes: Internal Anchor	Model	Reference
3. True scores on test form are linear function of true scores on anchor test.							
a. As stated above	Cross-Products Matrices	x_a, y_b	x_a, y_b	LXY-E[e]	LXY-I[f]	Lord X Y	Lord, 1976
b. As stated above	Cross-Products Matrices	v_a, v_b	v_a, v_b	LV-E[e]	LV-I[f]	Lord V	Lord, 1976
c. As stated above	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	LML-E[e]	LML-I[f]	Lord Maximum Likelihood	Lord, 1976; Birch, 1964
4. True scores on part test are linear function of true scores on anchor test.	Cross-Products Matrices	None	--	LCS	--	Lord Congeneric Subtests	Lord, 1976
IV. Curvilinear True Score Models							
A. X and V and Y and V Congeneric							
1. Item response function represented by a 1-parameter logistic function	Item Responses	None	None	ICC-1	ICC-1	1-Parameter Logistic ICC	Lord, 1975
2. Item response function represented by a 3-parameter logistic function	Item Responses	None	None	ICC-3	ICC-3	3-Parameter Logistic ICC	Lord, 1975

[a-f] These models yield identical conversion parameters when the Angoff method of estimating error variances is used.
[g-i] These models always yield identical conversion parameters.

Notation: x, y, v Observed score on new form, old form, and anchor test, respectively.
x', y', v' True scores.
a, b, c Group taking tests X and V; group taking tests Y and V; combined group.

Table 2

Design for Comparing Linear, Equipercentile, and ICC Equating Models

Equating a Medium-Difficulty Test to Itself through an Anchor Test of Similar Content

Test Variation	Anchor Test		Relation Between Samples	SAT-Verbal Score Level	
	Location	Difficulty		New Form Sample	Old Form Sample
1	External	Medium	Random	Middle	Middle
			Dissimilar	Middle	High
4	Internal	Medium	Random	Middle	Middle
			Dissimilar	Middle	High
5	Internal	Easy	Random	Middle	Middle
			Dissimilar	Middle	High
	Internal	Hard	Random	Middle	Middle
			Dissimilar	Middle	High

Equating Tests that Differ Only in Difficulty through an Internal Anchor Test of Similar Content and Medium Difficulty

Test Variation	Total Test Difficulty		Relation Between Samples	SAT-Verbal Score Level	
	New Form	Old Form		New Form Sample	Old Form Sample
8	Easy	Medium	Random	Middle	Middle
			Dissimilar	Low	Middle
	Medium	Hard	Random	Middle	Middle
			Dissimilar	Middle	High
9	Easy	Hard	Random	Middle	Middle
			Dissimilar	Low	High

Table 3

Item Summary Statistics and Identification Codes for Tests Used in Equating

Administration	Test Description	ID	n	$\bar{\Delta}_e$	σ_{Δ_e}	$\bar{r}_b$
	Equating a Test to Itself					
April 1975	Total Test:					
	SAT-Verbal	FT2XX	85	11.40	3.38	.47
	Anchor Test:					
	External	FE2FM	40	11.14	3.06	.43
	Internal					
	Easy	FE2DE	34	9.39	2.90	.49
	Medium	FE2DM	34	11.40	3.26	.49
	Hard	FE2DH	34	12.96	2.49	.47
November 1975	Total Test:					
	SAT-Verbal	FT4XX	85	11.36	3.40	.48[a]
	Anchor Test:					
	External	FE4FG	40	12.01	2.95	.43
	Internal					
	Easy	FE4DE	36	9.40	2.68	.51
	Medium	FE4DM	36	11.44	3.43	.49
	Hard	FE4DH	36	13.29	2.18	.48
	Equating Different Tests					
April 1975	Total Tests:					
	Easy	FT2DE	54	9.34	2.89	.46
	Medium	FT2DM	54	11.34	2.80	.46
	Hard	FT2DH	54	13.26	2.95	.44
	Anchor Test:					
	Internal	FE2PA	20	11.31	3.26	.46
November 1975	Total Tests:					
	Easy	FT4DE	54	9.59	2.94	.50
	Medium	FT4DM	54	11.69	2.49	.46
	Hard	FT4DH	54	13.51	2.77	.45
	Anchor Test:					
	Internal	FE4PA	20	11.58	3.05	.47

[a]Based on 84 of the 85 items.

Table 4

Formula Score Means, Standard Deviations, and Correlations[a] Between Anchor-Test and Total-Test Scores for Base Sample No. 32 (April 1975)

Test	No. of Items	Samples[b]				
		Random		Dissimilar		
		321XX	322XX	327XX	328XX	329XX
Equating a Test to Itself						
Total Test:						
FT2XX	85					
Mean		34.47	34.67		32.95	40.28
S.D.		15.48	15.57		14.47	14.82
Anchor Test:						
FE2DE	34					
Mean		19.91	19.96		19.43	22.27
S.D.		6.78	6.72		6.46	6.00
r_{xv}		.93	.93		.92	.92
FE2DM	34					
Mean		14.00	14.12		13.39	16.56
S.D.		6.98	6.98		6.58	6.66
r_{xv}		.94	.94		.93	.94
FE2DH	34					
Mean		9.15	9.27		8.45	11.83
S.D.		7.35	7.36		6.90	7.37
r_{xv}		.94	.94		.93	.95
FE2FM	40					
Mean		16.99	16.63		16.01	19.52
S.D.		8.06	8.06		7.78	7.86
r_{xv}		.87	.87		.85	.86
Equating a Test to a Different Test						
Total Test:						
FT4DE	54					
Mean		31.61		27.94		
S.D.		10.17		10.20		
r_{xv}		.88		.86		
FT4DM	54					
Mean		21.26	21.14		21.47	
S.D.		11.67	11.08		10.99	
r_{xv}		.90	.89		.89	
FT4DH	54					
Mean			11.89			16.17
S.D.			9.53			10.89
r_{xv}			.89			.89
Anchor Test:						
FE2PA	20					
Mean		8.46	8.42	7.25	8.02	9.72
S.D.		4.15	4.17	3.97	3.99	4.07

[a]Correlations are between the indicated anchor-test score and total-test score FT2XX in the case of a test being equated to itself, and between anchor-test score FE2PA and the indicated total-test score in the case of a test being equated to a different test. All anchor-test scores were included in the total-test score except for FE2FM.

[b]The last two digits of the sample number refer to the total-test and anchor-test score combination used for a particular equating.

Table 5

Formula Score Means, Standard Deviations, and Correlations[a] Between Anchor-Test and Total-Test Scores for Base Sample No. 44 (November 1975)

Test	No. of Items	Samples[b]				
		Random		Dissimilar		
		441XX	442XX	447XX	448XX	449XX
Equating a Test to Itself						
Total Test:						
FT4XX	85					
Mean		36.34	35.97		36.70	42.18
S.D.		15.77	14.92		14.74	15.10
Anchor Test:						
FE4DE	36					
Mean		21.82	21.74		22.15	24.19
S.D.		7.13	7.11		6.76	6.43
r_{xv}		.93	.93		.93	.92
FE4DM	36					
Mean		15.31	15.23		15.47	17.81
S.D.		7.13	6.85		6.75	6.89
r_{xv}		.95	.95		.94	.95
FE4DH	36					
Mean		9.19	8.98		9.12	12.02
S.D.		7.85	7.29		7.56	8.06
r_{xv}		.94	.93		.93	.95
FE4FG	40					
Mean		14.38	14.15		14.57	17.22
S.D.		8.32	8.02		7.80	8.24
r_{xv}		.87	.86		.85	.87
Equating a Test to a Different Test						
Total Test:						
FT2DE	54					
Mean		31.88		28.61		
S.D.		10.05		10.25		
r_{xv}		.89		0.87		
FT2DM	54					
Mean		21.99	22.07		20.85	
S.D.		10.95	10.92		10.38	
r_{xv}		.91	.90		.89	
FT2DH	54					
Mean			12.94			16.47
S.D.			10.21			10.45
r_{xv}			.89			.89
Anchor Test:						
FE4PA	20					
Mean		8.12	7.93	6.66	8.08	9.48
S.D.		4.36	4.13	3.89	4.07	4.39

[a]Correlations are between the indicated anchor-test score and total-test score FT4XX in the case of a test being equated to itself and between anchor-test score FE4PA and the indicated total-test score in the case of a test being equated to a different test. All anchor-test scores were included in the total-test score except for FE4FG.

[b]The last two digits of the sample number refer to the total-test and anchor-test score combination used for a particular equating.

Table 6

Mean Square Errors and Squared Biases for Comparing Equating Models

Where a Test is Equated to Itself through a Medium-Difficulty Equating Test (Test Variations 1 and 4)

Test Var.	Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
						Random Samples - Mean Square Error						
4	A.129	FT2XX	32102	FT2XX	32202	FE2DM	FE-T-A	.29	9.30	6.25	7.67	.56
1	A.137	FT2XX	32110	FT2XX	32210	FE2FM	PAC	2.34	63.20	45.02	26.83	18.40
4	A.236	FT4XX	44102	FT4XX	44202	FE4DM	LXY-IF	1.59	17.98	13.40	3.72	2.56
1	A.244	FT4XX	44110	FT4XX	44210	FE4FG	PB	1.74	17.89	16.56	4.54	1.06
						Dissimilar Samples - Mean Square Error						
4	A.184	FT2XX	32802	FT2XX	32902	FE2DM	LML-IF	.13	22.28	35.76	1.72	20.98
1	A.188	FT2XX	32810	FT2XX	32910	FE2FM	PB	4.88	58.06	146.65	17.14	8.53
4	A.291	FT4XX	44802	FT4XX	44902	FE4DM	T2O-IF	.18	15.44	25.20	2.19	.85
1	A.295	FT4XX	44810	FT4XX	44910	FE4FG	LV-IR1	16.65	33.18	77.44	39.31	2.66
						Random Samples - Squared Bias[b]						
4	A.129	FT2XX	32102	FT2XX	32202	FE2DM	FE-T-A	(-) .21	(-) .32	(-) .24	(-) 3.66	(-) .56
1	A.137	FT2XX	32110	FT2XX	32210	FE2FM	PAC	(+) 1.71	(+)34.92	(+) 28.16	(+)25.76	(+)18.22
4	A.236	FT4XX	44102	FT4XX	44202	FE4DM	LXY-IF	(-) 1.59	(-) 2.01	(-) 1.08	(-) 3.45	(-) 2.51
1	A.244	FT4XX	44110	FT4XX	44210	FE4FG	PB	(+) .65	(+) .24	(+) .32	(-) 2.38	(-) .39
						Dissimilar Samples - Squared Bias[b]						
4	A.184	FT2XX	32802	FT2XX	32902	FE2DM	LML-IF	(-) .01	(+) 3.02	(+) 22.00	(-) .06	(-)19.51
1	A.188	FT2XX	32810	FT2XX	32910	FE2FM	PB	(-) 4.83	(+)24.08	(+)133.29	(+)11.14	(+) 4.03
4	A.291	FT4XX	44802	FT4XX	44902	FE4DM	T2O-IF	(+) .18	(+) 5.82	(+) 18.83	(+) 1.94	(-) .15
1	A.295	FT4XX	44810	FT4XX	44910	FE4FG	LV-IR1	(+)10.48	(+)18.65	(+) 69.88	(+)20.36	(+) 2.53

[a]Scores were based on the following numbers of items: FT2XX and FT4XX: 85; FE2FM and FE4FG: 40; FE2DM: 34; FE4DM: 36.

[b]The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 7

Mean Square Errors and Squared Biases for Comparing Equating Models

Where a Test is Equated to Itself through an Easy or a Hard Equating Test (Test Variation 5)

Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
Random Samples - Mean Square Error											
A.127	FT2XX	32101	FT2XX	32201	FE2DE	LXY-IR1	.22	11.90	6.05	3.28	.34
A.131	FT2XX	32103	FT2XX	32203	FE2DH	T1	.31	10.11	6.15	8.76	.20
A.234	FT4XX	44101	FT4XX	44201	FE4DE	PAC	6.10	33.87	27.98	11.63	1.37
A.238	FT4XX	44103	FT4XX	44203	FE4DH	FE-T-F	.45	10.18	8.35	5.20	.11
Dissimilar Samples - Mean Square Error											
A.183	FT2XX	32801	FT2XX	32901	FE2DE	LXY-IF	30.14	29.70	53.73	2.34	2.69
A.185	FT2XX	32803	FT2XX	32903	FE2DH	LV-IR1	22.85	17.39	39.94	1.28	2.07
A.290	FT4XX	44801	FT4XX	44901	FE4DE	FE-T-A	59.14	36.36	61.94	15.37	2.40
A.292	FT4XX	44803	FT4XX	44903	FE4DH	FE-O	1.19	19.10	9.18	2.72	5.34
Random Samples - Squared Bias[b]											
A.127	FT2XX	32101	FT2XX	32201	FE2DE	LXY-IR1	(+) .21	(+) .31	(+) .40	(+) .26	(+) .27
A.131	FT2XX	32103	FT2XX	32203	FE2DH	T1	(-) .08	(-) .41	(-) .21	(-) 3.63	(-) .16
A.234	FT4XX	44101	FT4XX	44201	FE4DE	PAC	(-) 6.07	(-) 1.36	(-) .88	(-) 2.53	(-) .82
A.238	FT4XX	44103	FT4XX	44203	FE4DH	FE-T-F	(-) .45	(-) .64	(-) .37	(-) 1.07	(-) .06
Dissimilar Samples - Squared Bias[b]											
A.183	FT2XX	32801	FT2XX	32901	FE2DE	LXY-IF	(-) 1.27	(+)11.66	(+)41.91	(+) .17	(-) .65
A.185	FT2XX	32803	FT2XX	32903	FE2DH	LV-IR1	(+) 3.04	(+) 5.44	(+)28.10	(+) .60	(-) .59
A.290	FT4XX	44801	FT4XX	44901	FE4DE	FE-T-A	(+)58.19	(+)27.93	(+)56.22	(+)14.70	(+)2.33
A.292	FT4XX	44803	FT4XX	44903	FE4DH	FE-O	(+) .90	(-) 1.78	(+) .87	(-) 1.16	(-)4.88

[a]Scores were based on the following numbers of items: FT2XX and FT4XX: 85; FE2FM and FE4FG: 40; FE2DM: 34; FE4DM: 36.

[b]The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 8

Mean Square Errors and Squared Biases (ICC Equipercentile Criterion) for Comparing Equating Models Where a Test is Equated to a Different, More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 8)

Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
Random Samples - Mean Square Error											
A.196	FT2DE	32115	FT2DM	32217	FE2PA	TML-IA[b]	411.28	27.77	15.37	10.82	72.93
A.198	FT2DM	32117	FT2DH	32216	FE2PA	LV-IF	547.56	33.18	23.91	22.94	225.90
A.303	FT4DE	44115	FT4DM	44217	FE4PA	FE-O	656.90	38.94	40.32	45.97	169.00
A.305	FT4DM	44117	FT4DH	44216	FE4PA	LV-IF	469.59	71.23	71.57	37.09	207.94
Dissimilar Samples - Mean Square Error											
A.224	FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	417.79	25.50	31.36	15.84	68.89
A.232	FT2DM	32817	FT2DH	32916	FE2PA	FE-T-A	608.61	59.14	105.06	4.49	211.12
A.331	FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	679.12	77.44	116.21	35.40	160.53
A.339	FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	482.68	40.07	64.80	11.36	249.32
Random Samples - Squared Bias[c]											
A.196	FT2DE	32115	FT2DM	32217	FE2PA	TML-IA[b]	(+) 1.47	(+) ..77	(+) .37	(-) 5.79	(+) .02
A.198	FT2DM	32117	FT2DH	32216	FE2PA	LV-IF	(-) 1.85	(-) 3.78	(-) 3.86	(-)10.95	(-)107.26
A.303	FT4DE	44115	FT4DM	44217	FE4PA	FE-O	(+) 8.43	(+)12.08	(+) 9.49	(+)38.10	(+) 2.21
A.305	FT4DM	44117	FT4DH	44216	FE4PA	LV-IF	(-) 4.38	(-) 3.24	(-) 5.34	(-) 3.85	(-) 43.03
Dissimilar Samples - Squared Bias[c]											
A.224	FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	(-) .36	(+) 5.36	(+)17.27	(-)11.69	(+) 2.14
A.232	FT2DM	32817	FT2DH	32916	FE2PA	FE-T-A	(+)45.20	(+) 6.74	(+)50.63	(-) 2.29	(-) 93.00
A.331	FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	(-) 6.11	(+)24.56	(+)68.33	(+)22.42	(+) 4.41
A.339	FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	(-)13.45	(+) 1.06	(+)15.01	(-) 6.29	(-)101.20

[a]Scores were based on the following numbers of items: FT2DE, FT2DM, FT2DH, FT4DE, FT4DM, and FT4DH: 54; FE2PA and FE4PA: 20.

[b]Also TML-IF, TML-IR1, LUR-IA, LXY-IA, LV-IA, LML-IA, and PB.

[c]The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 9

Mean Square Errors and Squared Biases (Direct Equipercentile Criterion) for Comparing Equating Models Where a Test is Equated to a Different, More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 8)

Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
					Random Samples - Mean Square Error						
A.468	FT2DE	32115	FT2DM	32217	FE2PA	LML-IF	337.82	13.54	7.45	23.04	64.48
A.469	FT2DM	32117	FT2DH	32216	FE2PA	LV-IR1	361.00	8.01	9.92	34.57	102.62
A.473	FT4DE	44115	FT4DM	44217	FE4PA	FE-O	528.08	21.72	15.76	48.72	129.73
A.475	FT4DM	44117	FT4DH	44216	FE4PA	PD	273.90	13.54	11.90	36.36	100.00
					Dissimilar Samples - Mean Square Error						
A.470	FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	344.10	15.92	20.52	29.81	58.06
A.472	FT2DM	32817	FT2DH	32916	FE2PA	LV-IR1	431.39	43.30	114.28	19.62	89.11
A.476	FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	543.82	45.83	85.01	31.47	118.81
A.478	FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	275.23	25.10	44.22	34.57	129.73
					Random Samples - Squared Bias[b]						
A.468	FT2DE	32115	FT2DM	32217	FE2PA	LML-IF	(+) .80	(+) .28	(+) .07	(-) 7.68	(-) .05
A.469	FT2DM	32117	FT2DH	32216	FE2PA	LV-IR1	(+)2.53	(+) 1.06	(+) 1.02	(-) .11	(-)54.97
A.473	FT4DE	44115	FT4DM	44217	FE4PA	FE-O	(+)8.99	(+)12.84	(+) 10.12	(+)39.86	(+) 2.47
A.475	FT4DM	44117	FT4DH	44216	FE4PA	PD	(-) .01	(+) .36	(+) .01	(+) .20	(-)16.88
					Dissimilar Samples - Squared Bias[b]						
A.470	FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	(-) .91	(+) 3.92	(+) 14.72	(-)14.36	(+) 1.26
A.472	FT2DM	32817	FT2DH	32916	FE2PA	LV-IR1	(-)2.59	(+)31.30	(+)102.63	(+) 2.15	(-)44.89
A.476	FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	(-)5.97	(+)25.78	(+) 71.08	(+)23.59	(+) 4.77
A.478	FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	(-)1.56	(+)11.58	(+) 38.65	(-) .01	(-)57.31

[a]Scores were based on the following numbers of items: FT2DE, FT2DM, FT2DH, FT4DE, FT4DM, and FT4DH: 54; FE2PA and FE4PA: 20.

[b]The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 10

Mean Square Errors and Squared Biases (ICC Equipercentile Criterion) for Comparing Equating Models Where a Test is Equated to a Different, Much More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 9)

Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Models: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
Random Samples - Mean Square Error											
A.195	FT2DE	32115	FT2DH	32216	FE2PA	PAC	1,710.65	84.46	70.22	61.47	412.09
A.302	FT4DE	44115	FT4DH	44216	FE4PA	LXY-IR1	1,921.95	131.56	137.36	23.23	547.09
Dissimilar Samples - Mean Square Error											
A.225	FT2DE	32715	FT2DH	32916	FE2PA	PD	1,704.04	88.55	178.49	30.80	327.61
A.332	FT4DE	44715	FT4DH	44916	FE4PA	PD	1,927.21	166.15	241.80	30.58	495.95
Random Samples - Squared Bias[b]											
A.195	FT2DE	32115	FT2DH	32216	FE2PA	PAC	(-)12.94	(-)12.39	(-) 13.41	(-)43.24	(-)188.18
A.302	FT4DE	44115	FT4DH	44216	FE4PA	LXY-IR1	(-) 3.17	(-) 5.28	(-) 7.55	(-) .43	(-)137.14
Dissimilar Samples - Squared Bias[b]											
A.225	FT2DE	32715	FT2DH	32916	FE2PA	PD	(+) .37	(+)17.46	(+)101.59	(-)24.54	(-) 97.89
A.332	FT4DE	44715	FT4DH	44916	FE4PA	PD	(+) .38	(+)19.12	(+)117.57	(+) .85	(-) 94.08

[a]Scores were based on the following numbers of items: FT2DE, FT2DM, FT4DE, and FT4DH: 54; FE2PA and FE4PM: 20.

[b]The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 11

Mean Square Errors and Squared Biases (Direct Equipercentile Criterion) for Comparing Equating Models Where a Test is Equated to a Different, Much More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 9)

Ref. Table	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
					Random Samples - Mean Square Error						
A.467	FT2DE	32115	FT2DH	32216	FE2PA	PAC	1,253.16	10.76	12.89	101.40	247.43
A.474	FT4DE	44115	FT4DH	44216	FE4PA	LXY-IF	1,326.42	16.89	16.08	87.24	302.76
					Dissimilar Samples - Mean Square Error						
A.471	FT2DE	32715	FT2DH	32916	FE2PA	TML-IA[b]	1,254.58	69.39	191.27	49.98	171.61
A.477	FT4DE	44715	FT4DH	44916	FE4PA	T2O-IR1	1,334.44	85.75	215.50	53.29	254.40
					Random Samples - Squared Bias[c]						
A.467	FT2DE	32115	FT2DH	32216	FE2PA	PAC	(-) .15	(-) .09	(-) .19	(-)11.71	(-)114.61
A.474	FT4DE	44115	FT4DH	44216	FE4PA	LXY-IF	(+) 1.42	(+) .43	(+) .04	(+)13.37	(-) 79.15
					Dissimilar Samples - Squared Bias[c]						
A.471	FT2DE	32715	FT2DH	32916	FE2PA	TML-IA[b]	(-)24.93	(+)57.04	(+)184.43	(-) 3.11	(-) 46.30
A.477	FT4DE	44715	FT4DH	44916	FE4PA	T2O-IR1	(-) 7.53	(+)55.03	(+)195.79	(+)15.41	(-) 47.08

[a]Scores were based on the following numbers of items: FT2DE, FT2DH, FT4DE, and FT4DM: 54; FE2PA and FE4PA: 20.

[b]Also TML-IF, TML-IR1, LUR-IA, LXY-IA, LV-IA, LML-IA, and PB.

[c]The sign in parentheses inciates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

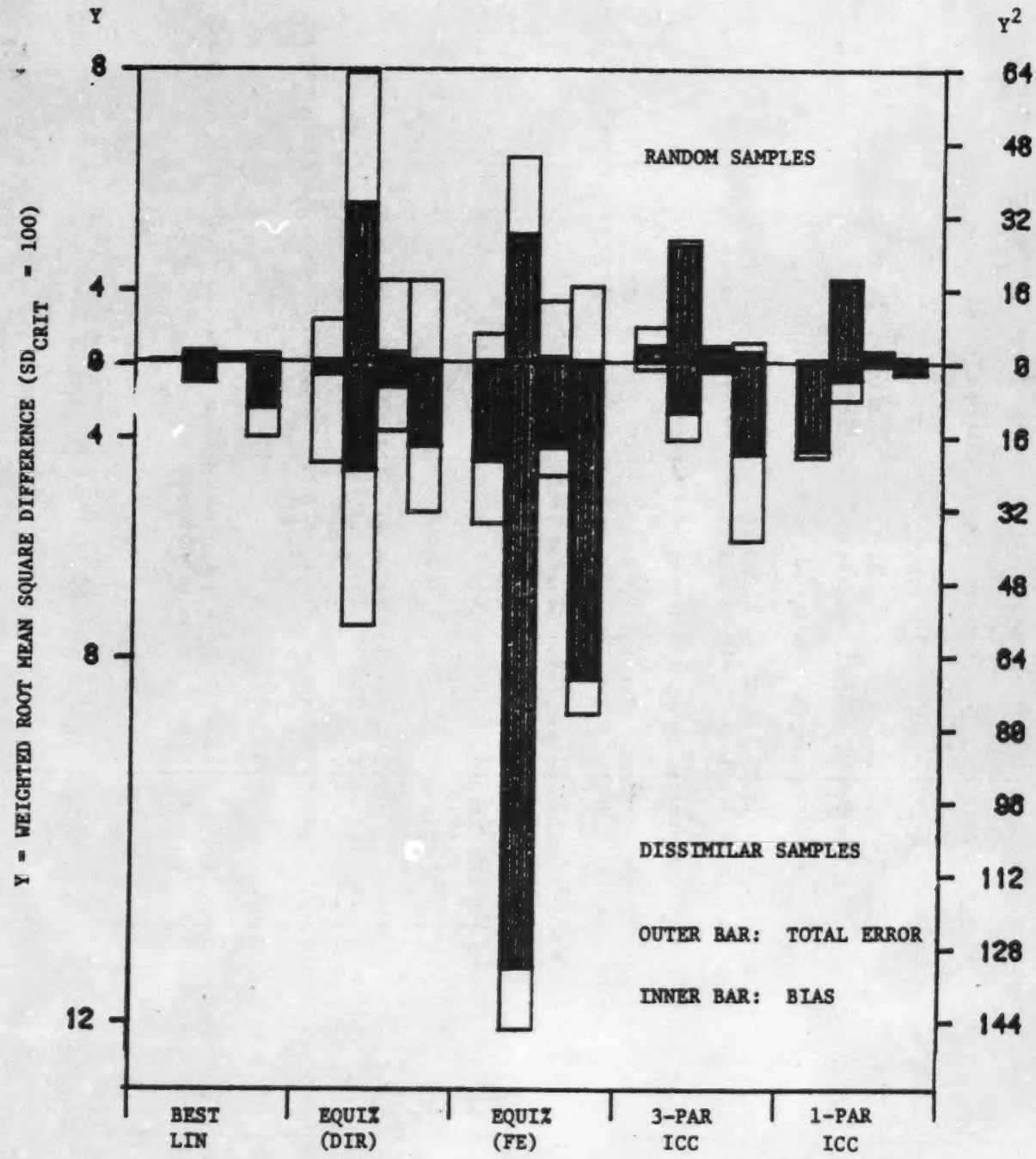

Figure 1. Comparisons of equating models: Equating a test to itself through a medium-difficulty anchor test.

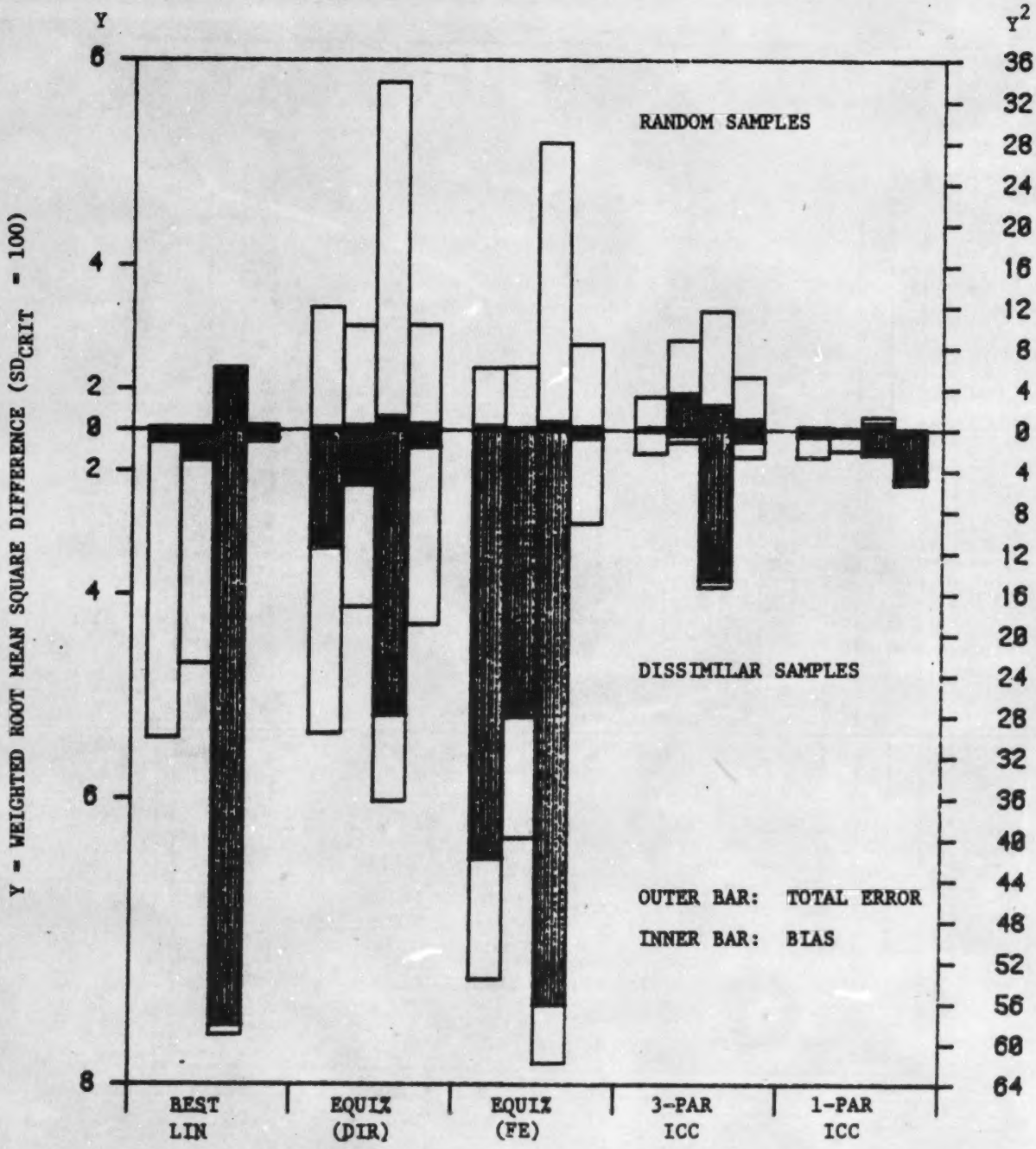

Figure 2. Comparisons of equating models: Equating a test to itself through an easy or hard anchor test.

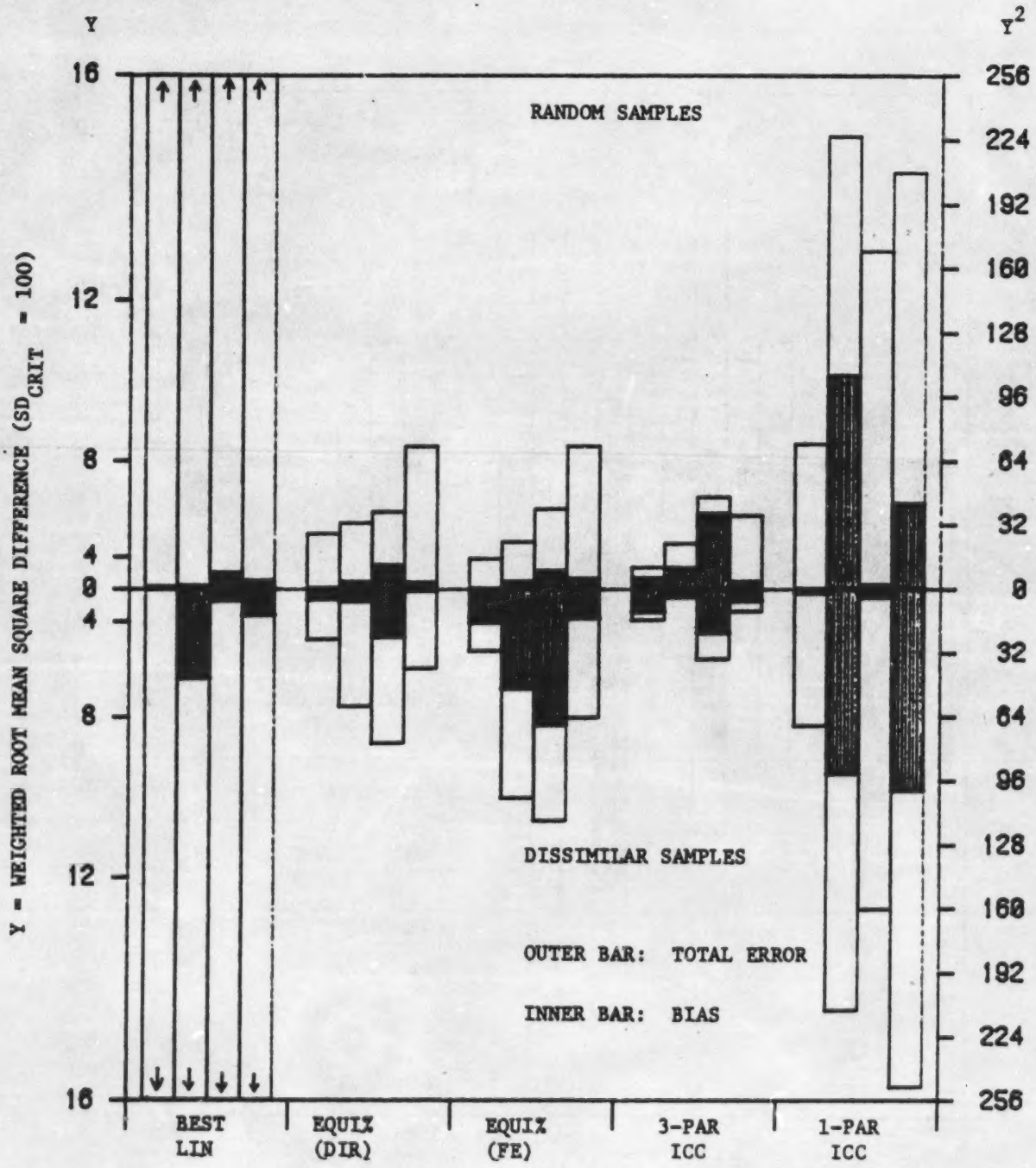

Figure 3. Comparisons of equating models: Equating a test to a different test using easy and medium or medium and hard total tests (ICC equipercentile criterion).

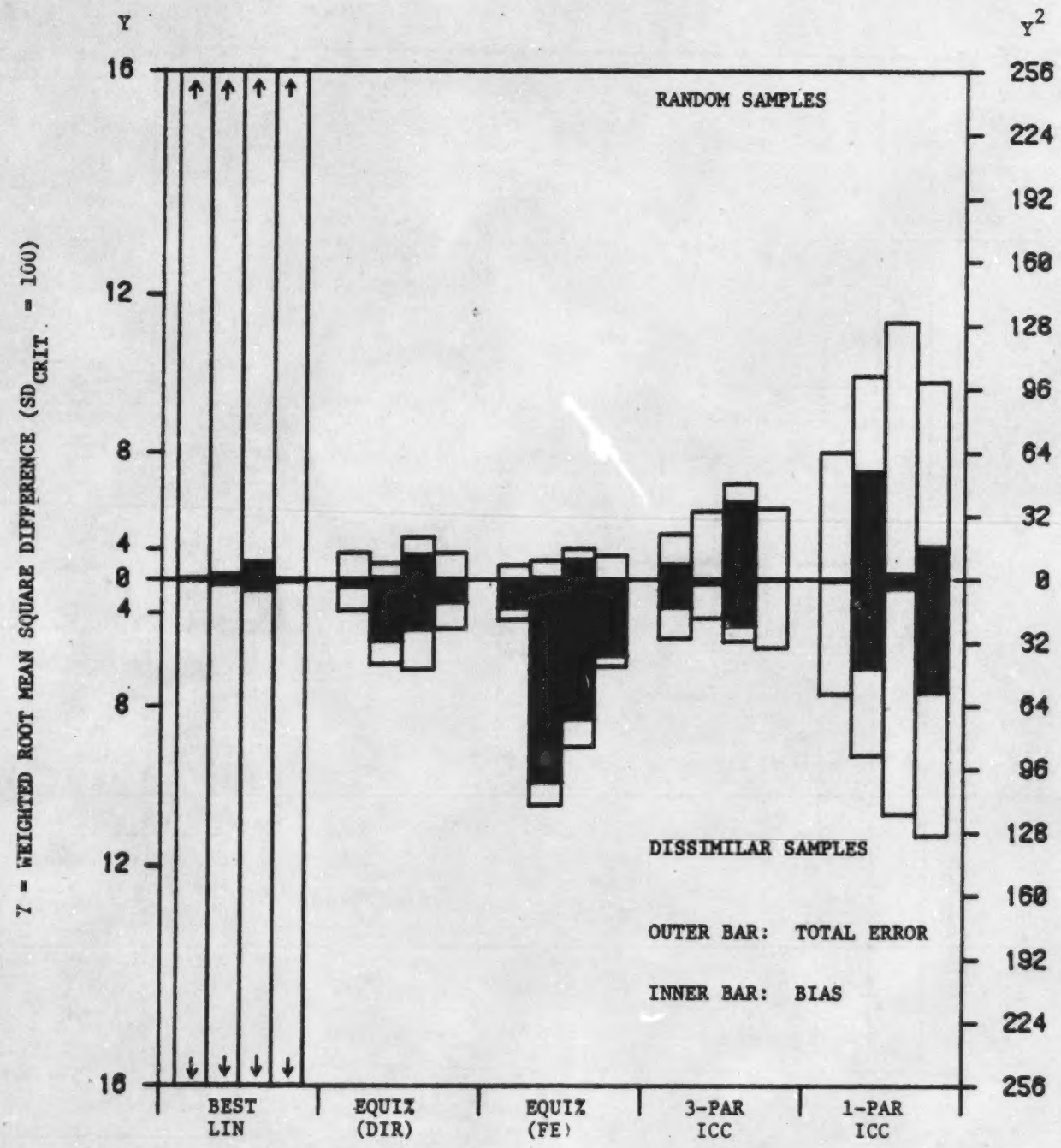

Figure 4. Comparisons of equating models: Equating a test to a different test using easy and medium or medium and hard total tests (direct equipercentile criterion).

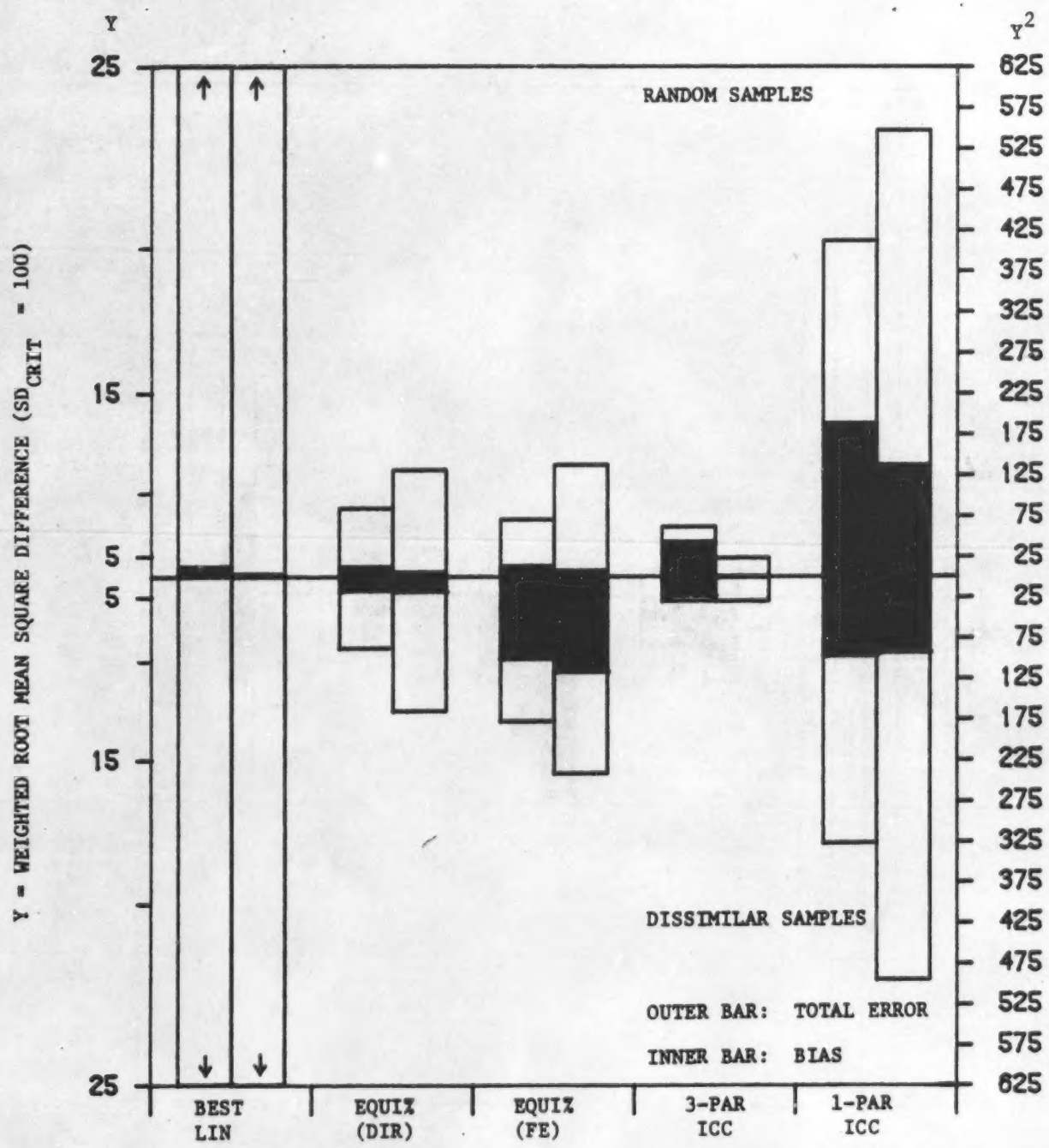

Figure 5. Comparisons of equating models: Equating a test to a different test using easy and hard total tests (ICC equipercentile criterion).

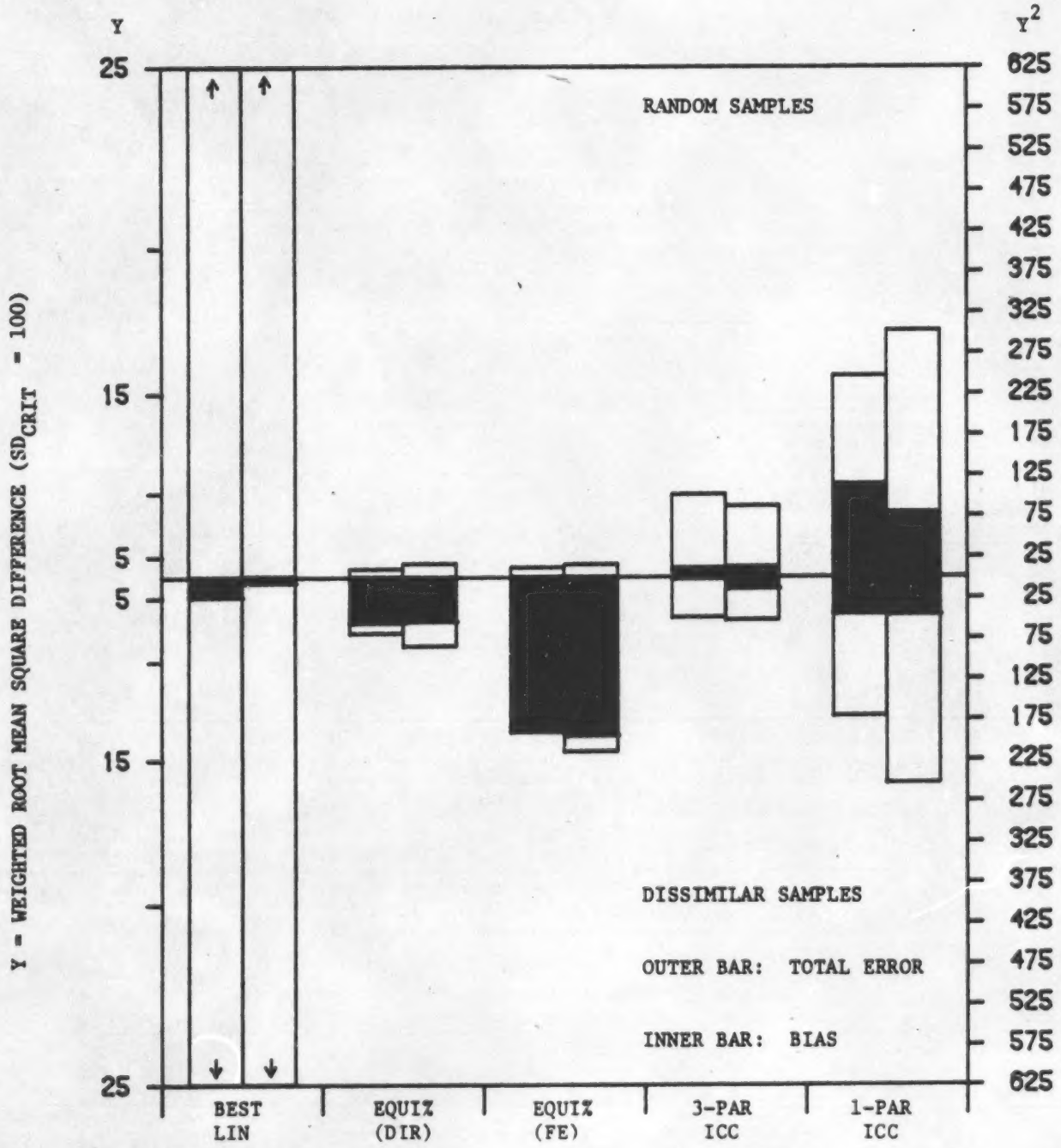

Figure 6. Comparisons of equating models: Equating a test to a different test using easy and hard total test (direct equipercentile criterion).